ONSET DETECTION EXPLOITING WAVELET TRANSFORM WITH BIDIRECTIONAL LONG SHORT-TERM MEMORY NEURAL NETWORKS

Giacomo Ferroni[2], **Erik Marchi**[1], **Florian Eyben**[1], **Stefano Squartini**[2], **Björn Schuller**[3,1]
[1]Machine Intelligence & Signal Processing Group, Technische Universität München, GERMANY
[2]A3LAB, Department of Information Engineering, Università Politecnica delle Marche, ITALY
[3]Department of Computing, Imperial College London, UK
{erik.marchi|eyben|schuller}@tum.de
giaferroni@gmail.com, s.squartini@univpm.it

ABSTRACT

A plethora of different onset detection methods have been proposed in the recent years. However few attempts have been made with regard to widely-applicable approaches in order to achieve superior performances over different types of music and with considerable temporal precision. This paper concerns the usage of Wavelet Packet Transform in order to exploits multi-resolution time-frequency features. We apply early fusion in the feature space by combining Wavelet Packet Energy Coefficients and auditory spectral features. The features are then processed by a bidirectional Long Short-Term Memory recurrent neural network, acting as reduction function. The network is trained with a large database of onset data covering various genres and onset types. Due to the data driven nature, our approach does not require the onset detection method and its parameters to be tuned to a particular type of music.

1. ALGORITHM DESCRIPTION

The algorithm can be seen divided in three parts. First, the audio data is transformed into the frequency domain via a Discrete Wavelet Packet Transform (DWPT) with 22 bands (cf. Table 1) and via two parallel STFTs with two different window sizes. Energy-based information and its evolution over time are used as the final feature set.

Second, the features are used as inputs to the BLSTM network, which produces an onset activation function as output.

Finally, the network output is post-processed by a thresholding and peak picking methods in order to obtain the correct position of the onsets. Figure 1 shows this procedure. The individual blocks are described in more detail in the following sections.

1.1 Feature Extraction

Discrete input audio files, sampled at $F_s = 44.1$kHz, have been used for our experiments.

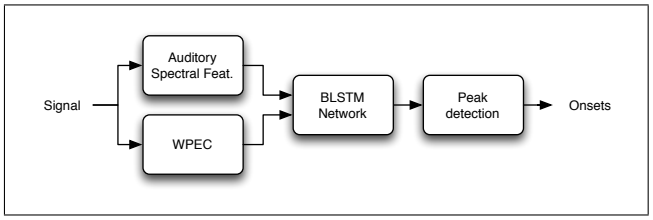

Figure 1. General block scheme.

A new features set is obtained exploiting wavelet transformation (cf. Figure 2) by obtaining Wavelet Packet Energy Coefficients (WPEC). The discrete input audio signal is segmented into overlapping frames of $W_{46} = 2048$ samples, which are sampled at a rate of 100 fps, log-energy of each frame is calculated before applying the Hamming window following:

$$E_n^{log} = \log\Big(\sum_k |x_n(k)|^2 + 1.0\Big) \qquad 1 \leq k \leq W_{46} \quad (1)$$

For each frame we computed the DWPT. By choosing b leaves of the decomposition-tree, we obtained b different representations of the original frame (one for each band).

Then, for each band, we calculated the energy $E_W(n,l)$, with n being the frame index and l the band index according to the following formula:

$$E_W(n,l) = \begin{cases} \sum_k (x_{n,l}[k])^2 + \sum_k (x_{n,l+1}[k])^2 & \text{if } l = 1 \\ \sum_k (x_{n,l-1}[k])^2 + \sum_k (x_{n,l}[k])^2 + & \\ \qquad + \sum_k (x_{n,l+1}[k])^2 & \text{if } 2 \leq l \leq b \\ \sum_k (x_{n,l-1}[k])^2 + \sum_k (x_{n,l}[k])^2 & \text{if } l = b \end{cases}$$

where $x_{n,l}[k]$ represents the DWPT coefficients corresponding to the l-th band and n-th frame. A logarithmic representation is chosen to match the human perception of loudness:

$$WPEC_{n,l}^b = \log(E_W(n,l) + 1.0) \qquad 1 \leq l \leq b \quad (2)$$

Finally the first order differences of the WPECs are calculated applying a half-wave rectifier function $H(x) = \frac{x+|x|}{2}$ to the difference of two WPECs one frame apart:

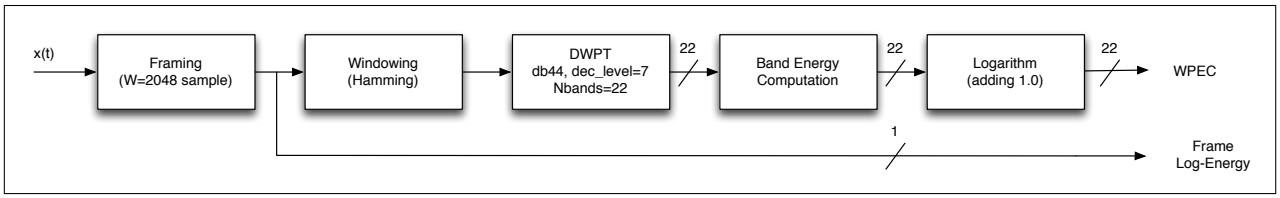

Figure 2. Wavelet packet energy coefficients extraction process. *db44* indicates the function used as mother wavelet (Daubechies of order 44), *dec_level* indicates maximum decomposition level needed and *Nbands* indicates the number of bands in which the signal is decomposed.

Level Bandwidth	N. Bands	Frequency Resolution
0 ÷ 2 kHz	12	172.27 Hz
2 ÷ 2.7 kHz	1	689.06 Hz
2.7 ÷ 11 kHz	6	1378.13 Hz
11 ÷ 16.5 kHz	2	2756.25 Hz
16.5 ÷ 22 kHz	1	5512.50 Hz

Table 1. DWPT frequency band division in detail.

$$WPEC^{+}_{n,l} = WPEC_{n,l} - WPEC_{n-2,l} \quad {}^{1\leq l\leq b} \quad (3)$$

In addition to WPECs we computed the well-know auditory spectral features [2]. The signal is divided into overlapping frames of W samples length ($W_{23} = 1024$ and $W_{46} = 2048$), that are sampled at the same rate of WPEC. The Hamming window is applied to these frames. For each window dimensions, the STFT gives the power spectrogram $S(n,k) = |X(n,k)|^2$, with n being the frame index, and k the frequency bin index. A conversion to the Mel-Frequency scale (by 40 triangular filter-bank) is made to reduce the dimensionality of STFT spectrogram. Furthermore a logarithmic representation is taken to mimic the human perception of loudness:

$$M_{log}(n,m) = \log(M(n,m) + 1.0) \quad (4)$$

We also applied the difference of two adjacent Mel spectrograms which leads to the positive first order differences $D^{+}(n,m)$, which carries information about the time evolution of the Mel-coefficients:

$$D^{+}(n,m) = M_{log}(n,m) - M_{log}(n-1,m) \quad (5)$$

1.2 BLSTM Neural Network

The best neural network for our purpose is a bidirectional RNN with LSTM units instead of usual non-linear units.

As network inputs we used 205 features per frame, composed in the following manner:

- 22 WPECs obtained (as in (2)) by the band division in Table 1 using the Daubechies wavelet function of order 44, 7-level of decomposition and $b = 22$ bands ($WPEC^{22}_{n,l}$).
- The log-energy of each frame (E^{log}_{n}) extracted as in (1).
- 22 WPEC positive differences ($WPEC^{+}_{n,l}$) as in (3).
- Two Mel-spectrograms ($M^{log}_{23}(n,m)$, $M^{log}_{46}(n,m)$) were computed with window size of 23.2 ms and 46.4 ms as in (4) and their corresponding first order positive differences ($D^{+}_{23}(n,m)$, $D^{+}_{46}(n,m)$), resulting in 160 features.

The network has six hidden layers in total (three for each direction) with 20 LSTM units each. The output layer has one unit and its output activation function lies between 0 and 1. It represents the probability for the class 'onset' and allows the use of the cross entropy error criterion to train the network [4].

1.2.1 Network Training and Dataset

Supervised learning with early stopping was applied to the network training. The dataset consists of 199 audio excerpts. It was created taking Bello's dataset [1], the dataset used by Glover et al. in [5], audio files used by Leveau et al. in [6] and some excerpts from ISMIR 2004 Ballroom set [1].

The final set was processed as monaural signals sampled at 44.1 kHz. It is composed by different categories of music [2] pitched percussive (PP e.g., piano), pitched non-percussive (PNP e.g., bowed strings), non-pitched percussive (NPP e.g., drums), complex mixture (MIX e.g., pop music) and others sound (OTHER is composed by ISMIR 2004 Ballroom dataset) for a total amount of 7989 onsets.

Presenting each audio sequence frame by frame to the network, its weights are recursively updated by standard gradient descent with backpropagation of the output error. The gradient descent algorithm requires the network weights to be initialised with non zero values. We initialise the weights with a random Gaussian distribution with mean 0 and standard deviation 0.1.

1.3 Peak Detection

The network obtained after training can classify each frame as 'onset' and consequently as 'non-onset' class. Frames

[1] `http://mtg.upf.edu/ismir2004/contest/tempoContest/node5.html`

[2] Bello and Glover datasets specify the music categories. ISMIR 2004 Ballroom dataset does not specify these information and we refer to it as OTHER.

containing the onsets are identified by processing the output unit function. Higher output activation function values indicate an high probability that the frame is an onset-frame.

An adaptive threshold technique has to be implemented before peak picking due to the dependency among the detection function, input signal, short time spectrum and wavelet packet coefficients.

In order to obtain the best classification for each song, a threshold θ is computed per song in accordance with the median and the mean of the activation function, fixing the range from $\theta_{min} = 0.1$ to $\theta_{max} = 0.3$:

$$\theta' = \lambda \cdot median\{a_0(1), ..., a_0(N)\} \quad (6)$$

$$\theta'' = \beta \cdot mean\{a_0(1), ..., a_0(N)\} \quad (7)$$

$$\theta = min(max(0.1, \theta', \theta''), 0.3) \quad (8)$$

where $a_o(n)$ is the output activation function of the BLSTM network (frames $n = 1...N$) and the scalar values λ and β are chosen to maximise the F-measure on the validation set. Their values are fixed to $\lambda = 50$ and $\beta = 3.7$.
The final onset function $o_o(n)$ contains only the activation values greater than this threshold.

$$o_o(n) = \begin{cases} 1 & o_o(n-1) \leq o_o(n) \geq o_o(n+1) \\ 0 & otherwise \end{cases}$$

2. RESULTS

The presented onset detector attained good performance in the MIREX 2013 evaluation (cf. gray row in Table 2).

Algorithm	F-measure	Precision	Recall
SB1	0.8727	0.8641	0.8946
ZHZD1	0.8233	0.7858	0.9009
FMESS1	0.8062	0.7770	0.8732
FMEGS1	0.8025	0.7880	0.8593
CF4	0.7345	0.6966	0.8507
CB1	0.6308	0.8506	0.5367
MTB1	0.3785	0.5429	0.3418

Table 2. Results for the MIREX 2013 onset detection evaluation. Only the best results of other participants or groups are shown with the exception for our two submissions.

3. ACKNOWLEDGMENT

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement No. 289021 (ASC-Inclusion).

4. REFERENCES

[1] J. P. Bello, L. Daudet, S. Abdallah, C. Duxbury, M. Davies and M. B. Sandler: "A tutorial on onset detection in music signals," *Speech and Audio Processing, IEEE Transactions*, 13.5:1035–1047, 2005.

[2] F. Eyben, S. Böck, B. Schuller and A. Graves: "Universal onset detection with bidirectional long short-term memory neural networks," *11th International Society for Music Information Retrieval Conference (ISMIR 2010*, 2010.

[3] S. Hochreiter and J. Schmidhuber: "Long short-term memory," *Neural Computation*, 9.8:1735-1780, 1997.

[4] A. Graves: "Supervised Sequence Labelling with Recurrent Neural Networks," *PhD thesis, Technische Universität München*. Munich, Germany. 2008.

[5] J. Glover, V. Lazzarini and J. Timoney: "Real-time detection of musical onsets with linear prediction and sinusoidal modeling," *EURASIP Journal on Advances in Signal Processing*, 1:1–13, 2011.

[6] P. Leveau, L. Daudet: "Methodology and tools for the evaluation of automatic onset detection algorithms in music," *In Proc. Int. Symp. on Music Information Retrieval*, 2004.